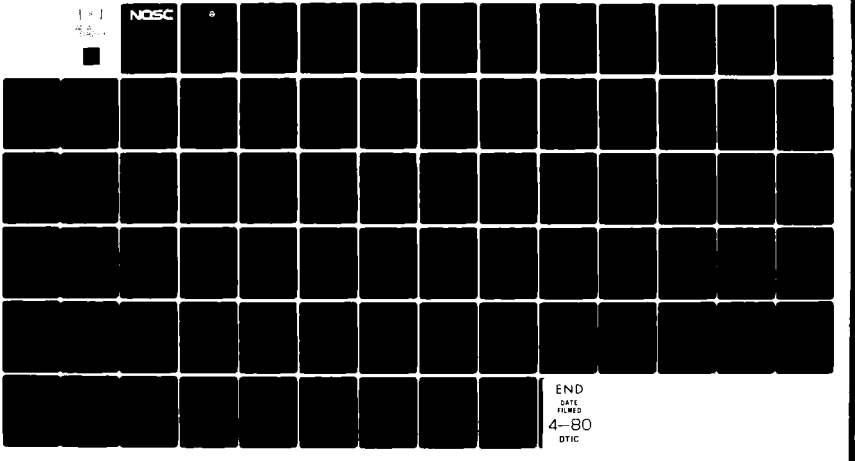
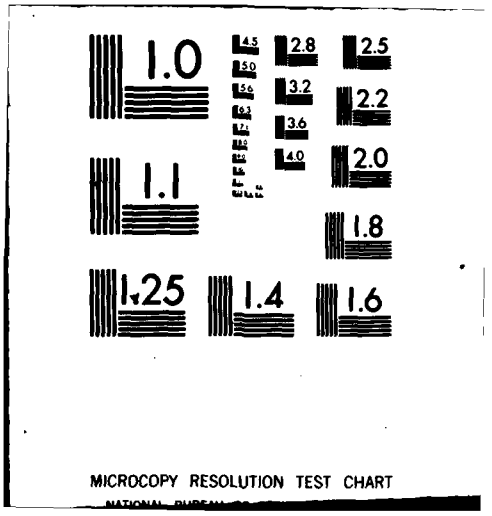
AD-A082 116 NAVAL OCEAN SYSTEMS CENTER SAN DIEGO CA F/G 20/1
ACOUSTIC MODEL EVALUATION PROCEDURES: A REVIEW. DISTANCE MEASUR--ETC(U)
SEP 79 R W MCGIRR
UNCLASSIFIED NOSC/TD-287 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
MICROCOPY RESOLUTION TEST CHART

NOSC

NOSC TD 287

Technical Document 287

ACOUSTIC MODEL EVALUATION PROCEDURES: A REVIEW

Distance measures and statistical test procedures for assessing the accuracy of propagation loss models

RW McGirr

September 1979

Prepared for
Naval Ocean Research and Development Activity
NSTL Station, Mississippi 39529

NAVAL OCEAN SYSTEMS CENTER
SAN DIEGO, CALIFORNIA 92152

NAVAL OCEAN SYSTEMS CENTER, SAN DIEGO, CA 92152

AN ACTIVITY OF THE NAVAL MATERIAL COMMAND

SL GUILLE, CAPT, USN
Commander

HL BLOOD
Technical Director

ADMINISTRATIVE INFORMATION

The work leading to this technical document was performed under NOSC Project R0120-SH, Program Element 63795N, "Acoustic Model Evaluation Committee Support," Principal Investigator: RW McGirr. The sponsoring activity is the Naval Ocean Research and Development Activity, Program Manager: Dr Aubrey Anderson. Funding for this effort was made possible jointly by the Acoustic Performance Prediction Program (PR Tiedeman, SEA 63D-5), the Surveillance Environmental Acoustic Support Program (LCDR K Evans, NORDA 600), and the Tactical ASW Environmental Acoustic Support Program (MG Lewis, NORDA 500) via the Acoustic Model Evaluation Committee (AMEC) under the auspices of OP-095E (R Winokur).

ACKNOWLEDGEMENT

The author wishes to express his appreciation to Dr Aubrey Anderson (NORDA), RB Lauer (NUSC/NLL), JD Pugh (NOSC) and LK Arndt (NOSC) for their support, technical review, and many helpful discussions.

Released by
MR Akers, Head
Systems Concepts and
Analysis Division

Under authority of
EB Tunstall, Head
Ocean Surveillance Systems Department

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NOSC Technical Document 287 (TD 287)
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): ACOUSTIC MODEL EVALUATION PROCEDURES: A REVIEW. Distance Measures and Statistical Test Procedures for Assessing the Accuracy of Propagation Loss Models.

5. TYPE OF REPORT & PERIOD COVERED:

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): RW McGirr

8. CONTRACT OR GRANT NUMBER(s):

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Ocean Systems Center, San Diego, CA 92152

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 63795N, R0120-SH

11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Ocean Research & Development Activity, NSTL Station, Mississippi 39529

12. REPORT DATE: September 1979

13. NUMBER OF PAGES: 74

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Acoustic model evaluation
Mathematical models
Mathematical analysis
Propagation models
Acoustic data
Model accuracy assessment

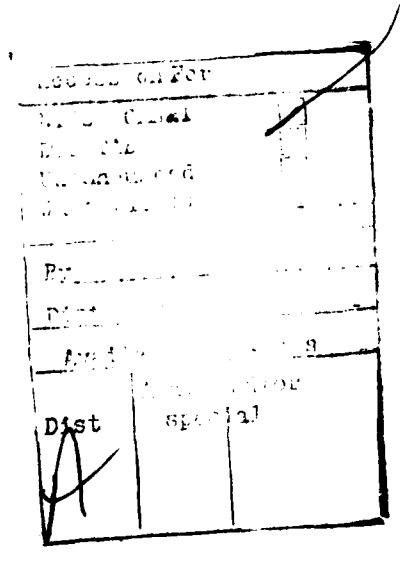
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The work presented in this document is a review of methods for evaluating computer models of underwater acoustic phenomena. More specifically, this work focuses on various approaches to the problem of quantitatively assessing the accuracy of propagation loss models. Model evaluation procedures recently formulated by two groups, the Panel on Sonar Systems Models (POSSM) and the Acoustic Environmental Support Detachment (AESD), are briefly reviewed. Various distance measures and statistical test procedures are explored as candidates for assessing the closeness of predictions and measurements.

DD FORM 1 JAN 73 1473 / EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

CONTENTS



1.0 INTRODUCTION

The performance of a sonar system in a specified ocean environment can be predicted for what occurs on the average, but not with great accuracy. There may be an untold number of reasons for prediction inaccuracies. The majority of such inaccuracies are usually attributed to the propagation loss model. As a consequence, evaluation of the predictive capability of propagation models (and acoustic models generally) is a topic of importance.

1.1 OBJECTIVE

The objective of this effort is to review acoustic model evaluation procedures. In an attempt to achieve greater clarity, attention is focused on the evaluation of propagation models. Most of the methods discussed here, however, are also applicable to reverberation models and, with some modifications, to ambient noise models.

1.2 BACKGROUND

The proliferation of sonar system performance prediction model development activity over the past decade is yielding a growing stockpile of models, some more accurate than others. As a consequence, those who use performance prediction models in analyzing competing design concepts are faced with the dilemma of selecting the "best" model. Unfortunately, assessing the merits of several candidate models, each programmed on a different computer, is a difficult task.

A related problem pertains to the validation of R&D models. Numerous propagation and ambient noise models have been developed under

6.2 funded programs. The rationale for this activity varies from case to case, but typically the requirement is based on either greater accuracy or higher computational speed. Whatever the reason, once a mathematical model has been transformed into a computer program it must then be subjected to test and evaluation procedures. Far too often the evaluation procedure employed by the model developer is too casual to engender high confidence in the model.

Many of the problems associated with model evaluation procedures can be rectified by community standardization. In an effort to achieve this standardization, the Acoustic Model Evaluation Committee (AMEC) was recently established. AMEC is comprised of representatives from Navy and university laboratories. It is chartered by OP-095E to establish a management structure and administrative procedures to evaluate environmental acoustic models of propagation, noise, and reverberation.

Prior to the conception of AMEC some work was accomplished in this endeavor by the Naval Underwater Systems Center (NUSC/NLL) under the guidance of the Panel On Sonar System Models (POSSM), and by the Acoustic Environmental Support Detachment (AESD) within the Model Evaluation Program (MEP). The procedures developed by POSSM and MEP are reviewed in section 3.

The POSSM/MEP review is preceded by a brief discussion of "conventional" evaluation methodology. This discussion provides background from which the unannointed reader should glean some appreciation of the problem. Following the POSSM/MEP review, several measures of

closeness suggested by procedures employed in hypothesis testing are discussed.

2.0 CONVENTIONAL EVALUATION METHODOLOGY

This section presents two aspects of conventional model evaluation methodology: (a) graphical comparisons and (b) insufficient replications. Their deficiencies are examined and exploited in subsequent sections as the rationale for evaluation strategies.

2.1 GRAPHICAL COMPARISONS

The procedure adopted by most model developers in assessing the accuracy of their product involves little more than overlaying plots of predictions on plots of measurements. Sometimes this procedure is extended by a brief examination of model response to variations in the controlling input parameters. More often than not, however, the model developer is so pleased with the obvious coincidence between predictions and measurements (within 10 dB all the way!) that further evaluation is deemed superfluous.

Graphical comparisons and sensitivity analyses are certainly important steps in the model evaluation process. They are especially useful in detecting gross departures of theory from experiment. Their usefulness and importance notwithstanding, these simple procedures lack the quantitative elements necessary to compare the performance of one model with that of another.

As demonstrated by figures 1 through 4 (from Forman [1975]), sole reliance on graphical comparisons does not always yield satisfactory conclusions. Curves generated by different models are superimposed on plots of measured data points. The graphical comparisons displayed in figure 1 are unambiguous in that the top curve obviously predicts better than the bottom curve. In this instance the two curves are separated by 8 or 9 dB, with one curve falling in the middle of the scattered data points and the other curve on the periphery of points. The situation depicted in figure 2 is less clear, however, because even though the proximity of curves to data points is similar to the previous case, the two curves are only 2 to 3 dB apart. Figure 3 illustrates curves generated by four different models. There is no question about which curve is the best predictor, but even the best curve displays a significant lack of fit. Of the three curves plotted in figure 4, the solid curve is obviously out of contention, but the predictive ability of the remaining two curves appears to be about the same, yielding an indeterminant circumstance.

---

1. Forman, L, Comparative Evaluation Methods for Propagation Loss Models, Computer Sciences Corporation Unpublished Report, Jul 1975.

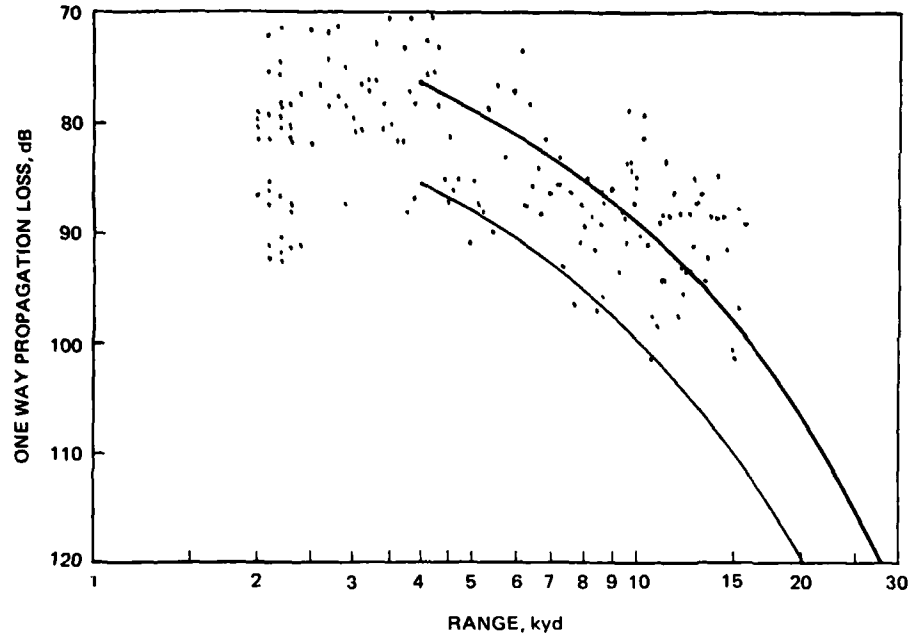


Figure 1. Unambiguous graphical comparison (from Forman [1975]).

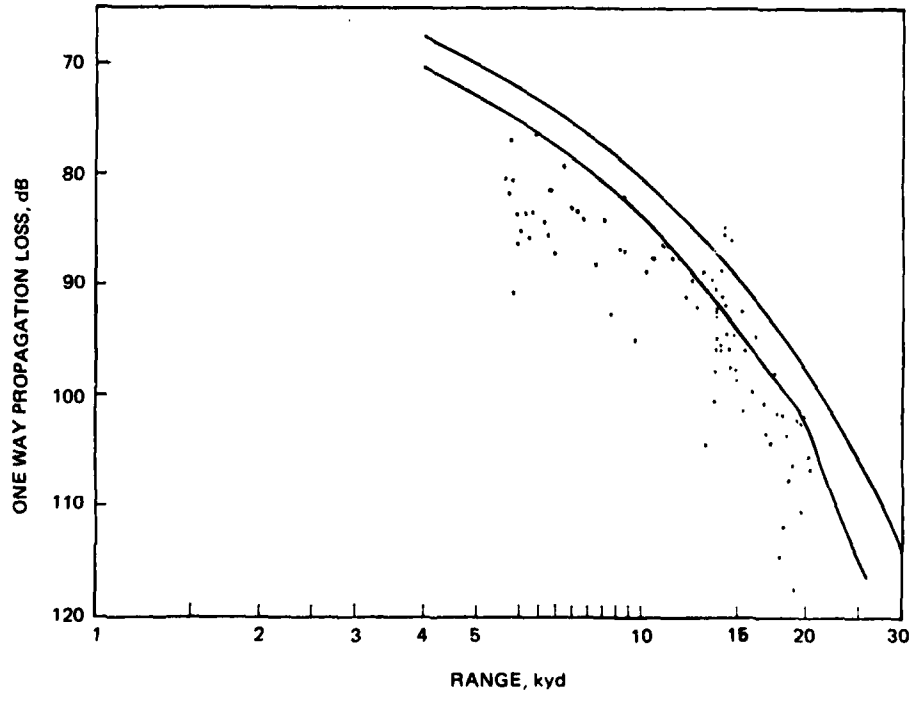


Figure 2. Ambiguous graphical comparison (from Forman [1975]).

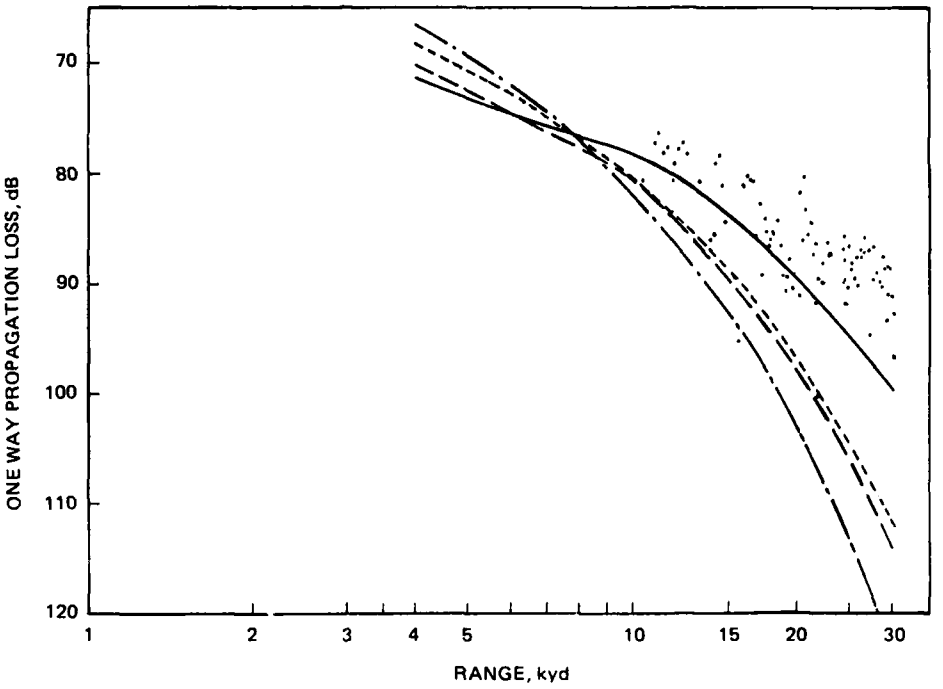


Figure 3. Qualified graphical comparison (from Forman [1975]).

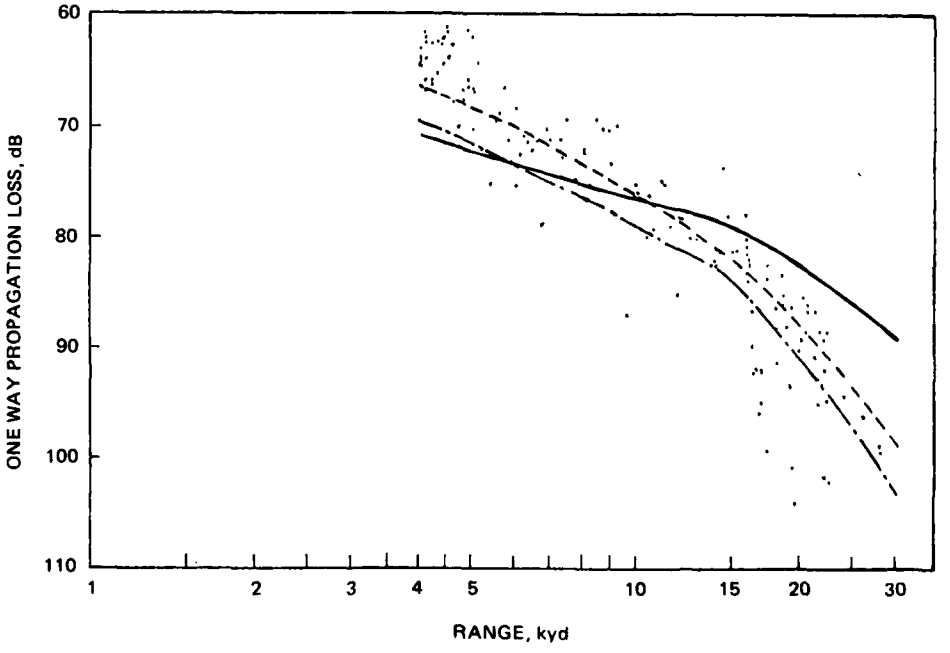


Figure 4. Inconclusive graphical comparison (from Forman [1975]).

2.2 INSUFFICIENT REPLICATIONS

To compound the problems of indeterminacy, data sets available for graphical comparisons far too often are insufficient in sample size, replications, and variety of conditions. The lack of replications probably represents the most frustrating deficiency among acoustic data sets. Apparently, at-sea measurement exercises are based on the assertion that independent measurements are unique. Moreover, experimental designs allowing for only two or three replications are apt to be as inconclusive as designs allowing for only one. Only when an experiment is replicated sufficiently is there likely to be a convergence of observed averages.

The question of what constitutes sufficient replication is beyond the scope of this discussion but, as an example of insufficient replication, consider the situation illustrated in figure 5. The two data sets were recorded simultaneously by two similar receiving arrays being towed away from a single source. The arrays were towed along parallel tracks within a few miles of each other, so that the propagation conditions were nearly identical. The dots represent data recorded on the SP LEE and the open circles represent data recorded on the USS BRONSTEIN. Of particular interest are the convergence zone (CZ) regions. The BRONSTEIN data have the leading edges at 29.5 nmi and 65 nmi, and the LEE data have them at 31 nmi and 62 nmi - an anomalous inversion. The relative amplitudes are also inverted from the first CZ to the second.

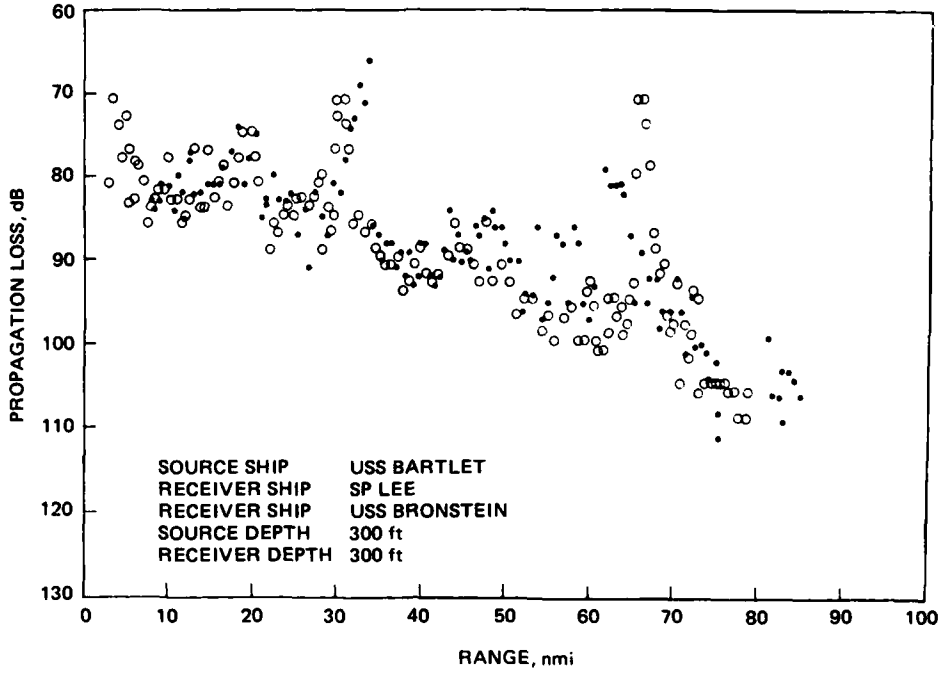


Figure 5. Example of "replicated" data sets.

Admittedly, these two data sets do not quite qualify as replications, but there is no evidence to suggest that the measurement conditions were substantially different. No matter what the sources of error or randomness, this example tends to illustrate that events do not recur precisely.

More to the point, however, only one sound velocity profile was obtained at the site, so that any predictions based on it would be expected to coincide equally well with both measured data sets. Such an expectation is contradictory unless the average of the two data sets is accepted as truth. Caution must be exercised in following such a practice, since the credibility of conclusions drawn from averages is more or less proportional to the number of data sets included.

2.3 CONSEQUENTIAL STRATEGIES

The indeterminate nature of graphical comparisons and the inadequate statistical base afforded by most acoustic data sets suggest two fundamental strategies. First, graphical comparisons should be complemented by quantitative measures of closeness. Second, similarities among data sets should be exploited to synthesize ensembles of (pseudo) replications.

3.0 POSSM AND MEP PROCEDURES

Two formal model evaluation efforts have emerged in recent years within the Navy's ocean acoustics community. One of these (POSSM) was a 6.2-funded effort sponsored by the Naval Sea Systems Command (O6H1), and the other (MEP) was a 6.3 funded effort carried out by the Acoustic Environmental Support Detachment (AESD). Some of their accomplishments are reviewed here.

3.1 POSSM PROCEDURES

The Panel on Sonar System Models (POSSM) was established in 1973 and chartered by the Sonar Technology Office (NAVSEA O6H1) to evaluate and make recommendations concerning propagation, ambient noise, and reverberation models to be used in NAVSEA sponsored sonar system programs. The philosophy adopted by POSSM immediately eclipsed the notion of recommending any single model to encompass the wide spectrum of potential tactical sonar applications. Instead, the model evaluation process would consist of assessing the attributes of candidate models and making these assessment findings available to NAVSEA users. The user could then select a model in accordance with requirements.

The membership of POSSM is comprised of system development engineers and acoustic model developers from the Navy laboratories. Model evaluation objectives and techniques are discussed at POSSM meetings, but most of the actual work involved in developing and implementing evaluation procedures has been accomplished at the Naval Underwater Systems Center (NUSC), New London Laboratory, under the direction of RB Lauer. The model accuracy assessment procedures adopted by POSSM

evolved from a methodology developed by Lauer and Skory [1975].[2] A brief account of POSSM model evaluation procedures is also presented in a technical report by DiNapoli [1975].[3]

In addition to model accuracy assessment, several other factors enter into model appraisal. Altogether the factors selected by POSSM include:

accuracy
execution time
storage requirements
implementation ease
execution complexity
program alteration ease
auxiliary outputs.

The first three factors lend themselves to quantitative description; whereas, the remaining factors elude precise definition and tend to accede to subjective specification. All of these factors are described

---

2. Lauer, RB and Skory, J, The Quantitative Comparison of Model Outputs with Experimental Data - A STAMP Program Application, Naval Underwater Systems Center TM TA11-46-75, 15 Jul 1975.
3. DiNapoli, FR, Computer Models for Underwater Sound Propagation, Naval Underwater Systems Center TD 5276, 31 Oct 1975.

in the "POSSM Reports" [Lauer and Sussman, 1976 and 1979][4,5] but brief descriptions of the last four factors are presented here for clarity.

Implementation ease relates to the level of difficulty involved in transforming a program from one machine to another and in becoming familiar with its operation.

Execution complexity pertains to input data requirements, especially as regards special control parameters peculiar to computational methods employed.

Program alteration ease bears on how well program documentation facilitates minor software modifications to improve efficiency or to accommodate special machine features.

Auxiliary outputs refers to outputs other than propagation loss versus range; for example, ray diagrams, travel time, or arrival angle structure.

4. Lauer, RB and Sussman, B, A Methodology for the Comparison of Models for Sonar Systems Applications, Volume I, Naval Sea Systems Command Report SEA 06H1/036-EVA/MOST-10, 9 Dec 1976.
5. Lauer, RB and Sussman, B, A Methodology for the Comparison of Models for Sonar System Applications, Volume II, Naval Sea Systems Command Report SEA06H1/036-EVA/MOST-11 (to be released).

One aspect of any model evaluation methodology that is likely to evoke controversy is the "objective" determination of model accuracy. The most straightforward approach entails forming residuals by taking the point-wise differences between measurements and predictions and then characterizing accuracy in terms of mean and standard deviations. This simple procedure has merit, but it also is susceptible to justified criticism. Much of the criticism stems from the stochastic character of measured data being somehow incompatible with the deterministi character of predictions. More specifically, second and higher order moments of measured data typically exhibit variation with range -- indicating range dependent distribution characteristics. As a consequence, a single, range-independent measure of accuracy is not likely to engender high confidence. On the other hand, the task of evaluating 20 or so models against several measured data sets or standards of comparison provides strong impetus to compress the number of accuracy measures to a minimum.

3.1.1 FIRST IMPLEMENTATION

In its first report [Lauer and Sussman, 1976][4] dealing with model evaluation methodology, POSSM attempts to treat this problem by comparing predictions to measurements (PARKA) in 20-kyd range subintervals over a total span of 100 kyds. The choice of 20-kyd intervals represents a compromise between two considerations: large enough to contain a sufficient number of sample points and small enough to constrain range-dependent features per subinterval to a minimum. Within each subinterval the mean and the standard deviations of residuals are

calculated. These results are then compressed into a single cumulative accuracy measure (CAM), defined as the sum of weighted averages of subinterval means and standard deviations. Mathematically CAM breaks down as

$$CAM = CAM\ (\mu) + CAM\ (\sigma),$$

where

$$CAM(\mu) = \sum_{n=1}^{N} \sum_{m=1}^{M_n} W_{nm} \mid \mu_{nm} \mid \Big/ \sum_{n=1}^{N} M_n$$

and

$$CAM(\sigma) = \sum\sum W'_{nm}\ \sigma_{nm} \Big/ \sum M_n$$

N is the number of cases compared against, M_n is the number of subintervals for each case, and the W_{nm} and W'_{nm} are weights that permit the relative importance of subintervals to be specified.

3.1.2 SECOND IMPLEMENTATION

A second volume similar to the first, but much more extensive in scope, [Lauer and Sussman, 1979][5] has been issued by POSSM. Twelve models are evaluated against Hays-Murphy Mediterranean Sea data. The evaluation procedures employed in volume I are again employed in volume II, with only minor differences. One of these differences, for example, is the way in which the total range interval (200 km) is divided into subintervals. Instead of equal length intervals, a model (RAYMODE X) is used to delineate direct path, bottom bounce, and convergence zone

boundaries out to the second bottom bounce region. Beyond this region the subintervals assume an arbitrary length of 50 km. The reason for using a model instead of the measured data to define subintervals is the lack of distinguishable structure in the data. Otherwise a model-based delineation scheme would be severely criticized.

The evaluation procedures that have evolved by virtue of POSSM represent the concerns of both model developers and model users. The two reports issued by POSSM (previously cited) reflect this evolutionary character as well as a certain amount of flexibility in procedure implementation. The format generally followed in both reports is presented in figure 6, and the steps followed in assessing model accuracy are as outlined in figure 7. Figures 8 through 10 illustrate how summary data pertaining to accuracy, speed, and storage requirements are presented (taken from volume II).

3.2 MEP PROCEDURES

The Model Evaluation Program (MEP) was initiated in 1973 by the Acoustic Environmental Support Detachment. Most of the work was performed during the 1973-1975 time frame, until AESD was moved to NSTL Station. Formal publications describing MEP procedures are not available, but accuracy assessment methods are discussed in a draft report by

- INTRODUCTION
- SCENARIO OF MEASURED DATA
- LIST OF CANDIDATE MODELS
- POINTS OF CONTACT
- BRIEF MODEL DESCRIPTIONS
    - ACCURACY ASSESSMENT
    - COMPUTER RUN TIMES
    - COMPUTER STORAGE REQUIREMENTS
    - COMPUTER FACILITIES
    - INPUT DATA REQUIRED
    - ANCILLARY OUTPUTS
        - SUMMARY AND CONCLUSIONS
        - REFERENCES
        - APPENDICES
            - > SPECIAL INPUT CONSIDERATIONS
            - > PL VS R PLOTS OF STANDARDS
            - > PL VS R PLOTS OF PREDICTIONS
            - > PLOTS OF RESIDUALS

Figure 6. *POSSM procedures, format of POSSM reports.*

1. SMOOTH COHERENT OUTPUTS
2. FORM RESIDUALS
3. SELECT SUBINTERVALS
4. CALCULATE SUBINTERVAL STATISTICS
5. ESTABLISH WEIGHTS
6. CALCULATE CAM(μ) AND CAM (σ)
7. CALCULATE CAM.

Figure 7. POSSM procedures,
steps followed in accuracy assessment.

Table 10A. Averages of Absolute Means and of Standard Deviations Over All Cases and Range Intervals (Standard of Comparison: Hays-Murphy Experimental Data).

	$\overline{\lvert\mu\rvert}$	$\bar{\sigma}$
AP2 NORMAL MODE	0.9	3.3
CONGRATS V COHERENT	2.4 (1.2)	4.5 (3.5)
CONGRATS V INCOHERENT	1.7 (0.9)	2.0 (1.8)
FACT/FNWC SEMI-COHERENT	1.2	2.0
FACT/NUSC COHERENT	1.1	2.5
FACT/NUSC SEMI-COHERENT	1.2	2.4
FACT/NUSC INCOHERENT	1.3	1.7
FFP 1/3-OCTAVE	3.6	2.6
GORDON NORMAL MODE	0.6[a]	3.4[a]
LORA COHERENT	1.9[b] (1.5)	4.7[b] (4.6)
LORA SEMI-COHERENT	1.4[b] (1.1)	2.5[b] (2.7)
LORA INCOHERENT	1.5[b] (1.0)	2.2[b] (2.2)
NSWC NORMAL MODE	1.2	4.0
PLRAY SEMI-COHERENT	1.1[c] (1.4)	2.3[c] (2.4)

(This reproduction has been abbreviated.)

Figure 8. (From Lauer and Sussman [1979]).

Table 13. Average Run Time Per Case.

Model (Computer)	Average Run Time Per Case	No. of Points Per Prediction (to 200 km)	Remarks
AP2 (CDC 6600)	60.6 sec	400	Run time is frequency dependent.
CONGRATS V (UNIVAC 1108)	42.2 sec	200	Includes both coherent and incoherent phase addition in each instance.
	70.0 sec	400	Resubmission.
FACT/FNWC (CDC 6500)	25 sec	216	Includes calculations other than those needed for propagation loss. Also, calculations were carried out for 250 points (125 nm).
FACT/NUSC (UNIVAC 1108)	2.5 sec	200	
FFP (CW) UNIVAC 1108)	6 min 13 sec	2372	Run time is frequency dependent. FFP (1/3 octave band average) run time is not included, since this is not a normal use of FFP. Calculations were carried out for 4096 points (345.2 km).

(This reproduction has been abbreviated.)

Figure 9. (From Lauer and Sussman [1979]).

Table 14. Storage Requirements.

		Storage Required			
Model	Computer	Instructions	Data	Total	Remarks
AP2	CDC 6600			20000	For 500 Modes
CONGRATS V	UNIVAC 1108	19000	21000	40000	Including Plot Routines
FACT/FNWC	CDC 6500			60000	"Estimated at 60000 Words Exclusive of Input/ Output Functions"
FACT/NUSC	UNIVAC 1108	14121	7734	21855	Without Plot Routines
FFP	UNIVAC 1108	18563	33009	51572	Without Plot Routine
GORDON NORMAL MODE	CDC 1110			53000	
LORA	UNIVAC 1108			38000	
NSWC NORMAL MODE	CDC 6500			18900	"45000 Octal Exclusive of Input/Output "
PLRAY	CDC 6600			20480	"Core Storage, 20480 Words (50,000 Octal). (This program has recently been revised to reduce the core requirements) "
RAYMODE X	UNIVAC 1108	2801 2912 3877 8536	3195 3320 3549 5579	5996 6232 7426 14115	Without I/O, w/o system With I/O, w/o system Without I/O, with system With I/O, with system
RAYWAVE II	UNIVAC 1110			37000	
RTRACE	CDC 3800			58100	"Including Input/Output and Library Functions "

(This reproduction has been abbreviated.)

Figure 10. (From Lauer and Sussman [1979]).

Cavanaugh [1974],[6] and software documentation is contained in a memorandum by Stieglitz [1974].[7]

Although there are many similarities between POSSM and MEP accuracy assessment procedures, there are two features of the MEP approach significantly distinctive to require separate review. The first feature is somewhat philosophical and pertains to a general analysis of errors. The second is a specific technique designed to analyze the range-dependence of errors. Both of these features are discussed in detail in the report by Cavanaugh [1974].[6]

3.2.1 ERROR ANALYSIS

The differences between measured and predicted loss form a sequence of observed errors $\{e_o\}_1^N$, there being N sample ranges. Each e_o in the sequence is treated as a random variable having a range-independent distribution. Moreover e_o is assumed to be the sum of two independent random variables, that is

$$e_o = e_m + e_p,$$

6. Cavanaugh, RC, Transmission Loss Model Evaluation Package, Part I: The Approach, Acoustic Environmental Support Detachment (unpublished report), 1 May 1974.
7. Stieglitz, R, Informal Documentation-Model Evaluation Package, Acoustic Environmental Support Detachment memo AESD:RS:dl of 28 Jun 1974.

where e_m represents measurement error and e_p represents prediction error. e_m can be broken down into

$$e_m = e_i + e_s,$$

where e_i represents error related to model input and e_s represents error associated with source level and processing. Each of these errors can be broken down further until all possible measurement parameters are accounted for. The level of error breakdown is dictated by the level of completeness of the measurements. Far too often, however, error bounds on such parameters as source depth, receiver depth, source level, and processing errors elude careful measurement, leaving only visceral confidence levels to rely on. In such circumstances an estimate of prediction error, e_p, can be obtained by resorting to the first-level formulation,

$$e_p = e_o - e_m,$$

where e_m is obtained from an ensemble of measured data sets.

These error analysis procedures are not difficult to implement, at least not conceptually. Unfortunately, not all measured data sets are complete enough to allow reasonable parameter estimations to be determined. Cavanaugh illustrates, by way of example, two approaches to the parameter estimation problem. In each case, however, a knowledge of error distributions is assumed. In this respect the error analysis methodology is incomplete, and further work in this area is desirable.

3.2.2 RANGE DISPLACEMENT ERROR

Cavanaugh introduces a procedure to account for errors associated with reported range values. The motivation for such a procedure derives from his error analysis development. However, the procedure has value independent of such a formulation. In particular, if two or more models are under evaluation and each exhibits a feature (e.g., CZ) displaced in range from that exhibited by measurement (see figure 11), then a quantitative measure of such displacement is desirable for each model.

The following interpretation of this procedure is reminiscent of "windowed" cross-correlation (see figure 12). For a given sample range, say R_n, the measured result is compared with predictions generated at several ranges within an interval covering R_n. The range displacement is then found for which the error is a minimum. This procedure is repeated for each sample range (or range interval). From the resulting set of range displacements, a histogram is generated depicting the frequency (or percent) of minima versus displacement. If the sample ranges are equispaced by, say, ΔR, then the set of range displacements consists of integers representing the number of ΔR bins displaced from the sample ranges. The range-displacement bin with the maximum frequency corresponds (more or less) to the lag-index of maximum cross correlation. If one bin has a frequency significantly greater than all other bins, then a translation bias is evident.

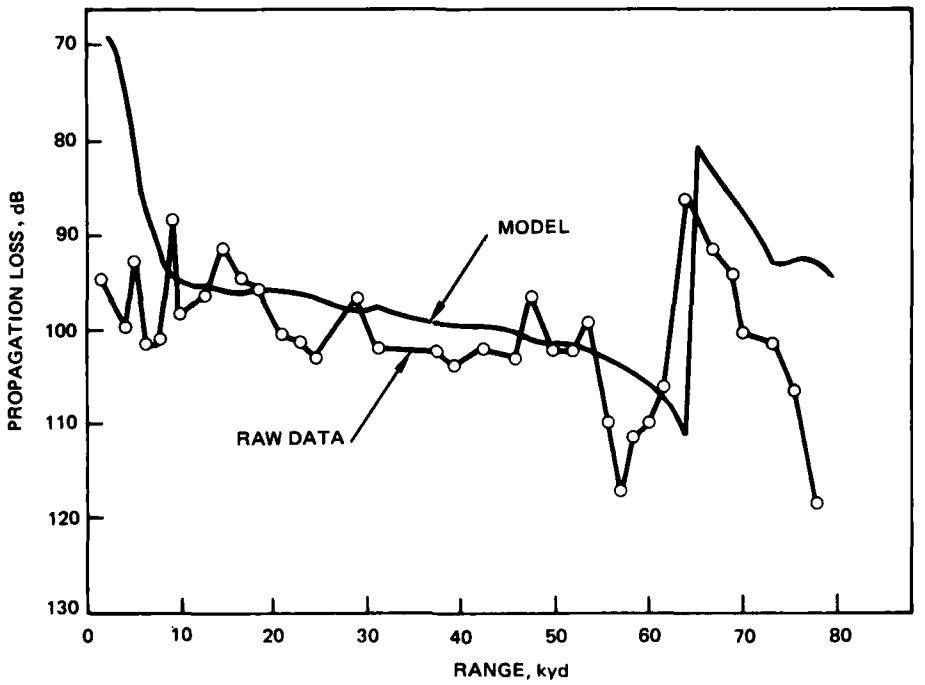


Figure 11. Translation invariance (from Forman [1975]).

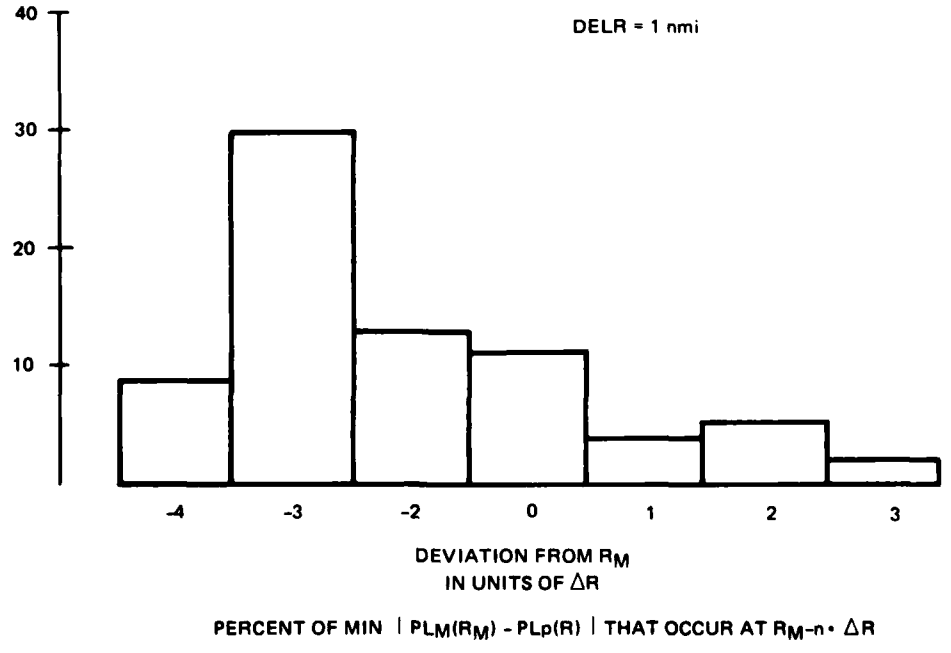


Figure 12. "Cross-correlation" histogram.

Actually, if a strong bias exists, it is readily apparent from visual inspection of superimposed plots. However, the range displacement error procedure provides a numerical estimate of the bias and allows model-to-model comparison.

Examples of some of the outputs available from the AESD Model Evaluation Package are presented in figures 13 through 15 [after Cavanaugh, 1974].[6]

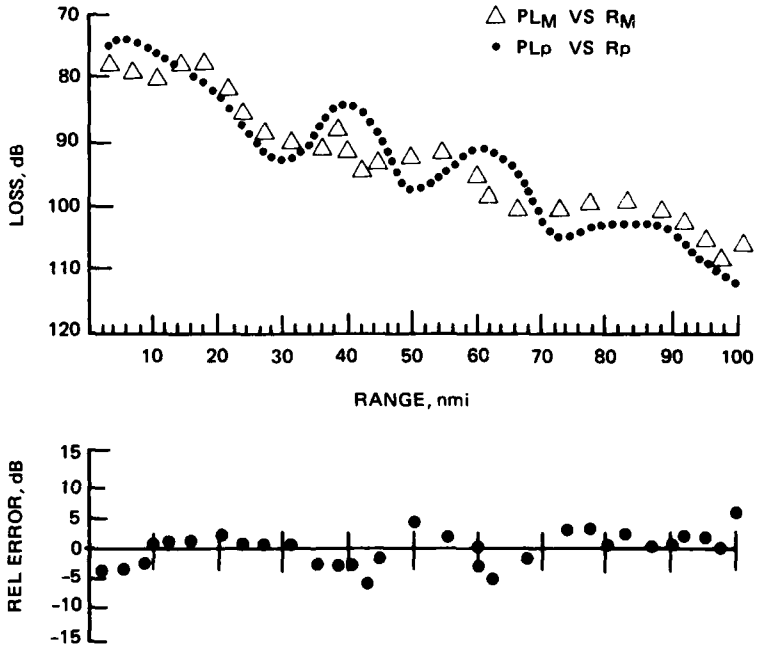


Figure 13. Standard comparative output.

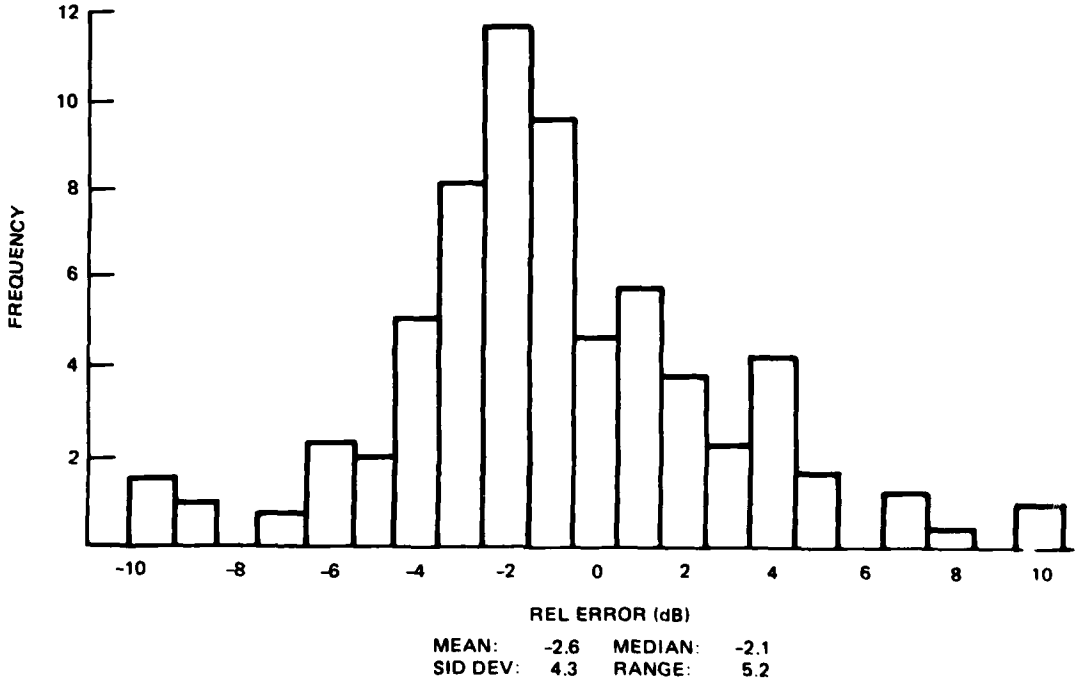


Figure 14. Number of errors vs error size (dB bins).

STATISTICS	INTERVAL NUMBER 1	2	3
SAMPLE SIZE	8	17	22
MEAN	-2.8	-1.5	1.2
STD DEV	3.1	3.6	4.2
SKEWNESS	-0.81	-0.42	-1.31
KURTOSIS	-0.93	-0.74	-0.98
ABS MEAN	3.1	2.4	2.8
ABS STD DEV	3.0	3.2	3.3

ENSEMBLE STATISTICS VERSUS RANGE INTERVAL

Figure 15. Range interval analysis.

4.0 MEASURES OF CLOSENESS

The typical approach taken in "goodness of fit" testing of linear regression models entails comparing a statistic derived from residual errors with a "critical" value. If the critical value is exceeded, the model is rejected as a good predictor of the observations. Higher-order models are successively tested in this manner until the derived statistic falls short of the critical value. Associated with the critical value is a parameter called the level of significance. Thus, a model that satisfies the test criteria does so at a pre-specified significance level.

The objectivity of this approach is appealing and is exploited in section 5. However, there are intuitive measures of closeness which deserve review irrespective of their subjective nature. The accuracy measures reviewed in this section derive from elements of classical statistics, real analysis, and the figure-of-merit approach to sonar systems analysis.

4.1 MEASURES SUGGESTED BY CLASSICAL STATISTICS

Quantities employed in classical analysis-of-variance (ANOVA) and regression procedures are exploited here as candidate measures of closeness. The utility of these quantities may be questionable under conditions which violate their underlying assumptions. However, the degree to which an assumption is violated can be determined, thereby providing a measure of credibility. As a minimum, most classical procedures require data samples to be independent and identically distributed. For example, the one-sample Student's t-test requires the

data to be normally distributed. Actually, a departure from normality is less cause for concern than either the lack of independence or a noticeably unstable variance. In some cases, the effects of contiguous correlations can be mitigated by means of decimation, thus reducing the original data set to a subset of "independent" samples. If range-dependent trends show up in the residual errors (or their mean or variance), the simplest procedure to circumvent trend effects entails little more than confining calculations to intervals of (relatively) constant variance.

4.1.1 REMARKS CONCERNING CALCULATIONS

More often than not, replicated data sets are not available, or conditions necessary to allow the assembly of independent data sets into ensembles are absent. In either case, the degrees of freedom desirable in calculating measures of closeness may well be less than optimal. Such circumstances necessitate the development of procedures appropriate for single-event data. An event, in this context, refers to data acquired along a single radial track, that is, a source closing toward or opening away from a stationary receiver at constant bearing.

Quantities calculated for both single-event and ensemble data are similar, and are discussed in the following sections. The idea behind these computations is to generate quantities that reflect the degree of closeness between predictions and measurements.

Quantities that can be routinely calculated as measures of closeness are based on residual errors obtained by taking the difference

between measurements and predictions. Let L_n denote measured propagation loss interpolated at sample range R_n, and let $\hat{L}_n$ denote the corresponding prediction. The residual error e_n, is

$$e_n = L_n - \hat{L}_n.$$

The sequence obtained by calculating a residual at each sample range R_n, n=1, 2, ..., N, forms the basis for the following sample statistics

$$\bar{e} = \sum e_n/N \qquad \text{mean}$$

$$S^2 = \sum (e_n - \bar{e})^2/(N-1) \qquad \text{variance}$$

$$d^2 = \sum (e_{n+1} - e_n)^2/(N-1) \qquad \text{mean square successive difference}$$

Measures of closeness commonly employed to assess the accuracy of linear regression models are:

(1) standard residual (RMS) error

$$S_e = \sqrt{\sum_n (L_n - \hat{L}_n)^2/(N-k)}$$

(2) correlation coefficient

$$\rho = \frac{\sum (L_n - \bar{L})(\hat{L}_n - \bar{L})}{\sqrt{\sum (L_n - \bar{L})(\hat{L}_n - \bar{L})}}$$

and

(3) ratio of variances

$$R^2 = \frac{\sum(\hat{L}_n - \bar{L})^2}{\sum(L_n - \bar{L})^2} \quad ,$$

where $\bar{L}$ is the mean value of measured data.

These expressions require slight modification for predictions generated by deterministic models vis-a-vis regression models. In the expression for S_e, N-k reflects the degrees of freedom, where k is the number of model coefficients estimated from data. For the class of deterministic models of interest here there are no "fit" coefficients, in which case k=0 is appropriate. The mean $\bar{L}$ used in the expressions for ρ and R^2 is the same for both predictions (via regression) and measurements as a consequence of the least squares criterion. However, for deterministic models there is no such criterion so that generally $\sum (L_n - \hat{L}_n) \neq 0$. There is temptation to simply replace $(\hat{L}_n - \bar{L})$ by $(\hat{L}_n - \hat{\bar{L}})$. Such a modification unfortunately affects the sensitivity of ρ significantly and reduces R^2 to nothing more than the ratio of prediction variation to measurement variation. Thus the expressions for ρ and R^2 may not be especially appropriate for acoustic model evaluation.

4.1.2 SUPPLEMENTAL MOMENTS

Global measures of closeness generated by single-event data are likely to indicate a significant lack of fit when there are distinct feature displacements. What is really needed is a "local" or range-

sensitive measure of closeness or several such measures. The need for a range-sensitive measure manifests itself in the features evident in most plots of propagation loss versus range. These features, representing departures from monotonicity, are subtle or even nonexistent for bottom limited situations, but are usually evident in RSR situations and in data processed coherently. Of immediate interest is the problem of assessing a model's ability to predict the location, width and magnitude of convergence zones. No single statistic or measure of closeness can perform this task. More than likely, a distinct measure is needed for each feature of interest.

Thus, as a matter of routine, global measures should be supplemented by sequentially-generated first and second order moments. These moments span only a few contiguous sample ranges, thereby serving as local measures of closeness. They also reveal the statistical nature of the residual errors as a function of range.

For data acquired under bottom limiting conditions, supplemental moments may be routinely calculated over subsamples of convenient size. The calculations assume a less routine character, however, when the measurement data abound in prominent features such as convergence zones. For such cases care is exercised in selecting subsamples to assure that no subsample is comprised of data propagated via both bottom reflected and deep refracted paths. This constraint is applied to both measured and predicted data to avoid excessively large residuals resulting from feature nonalignment. The subjective nature of this procedure potentially accedes to the unfortunate circumstance that the resulting

measures of closeness obtained for one model will not necessarily be commensurate with those obtained for another model.

Both global and sequentially generated moments can be used to generate confidence limits or bands about the mean residual error. If the variance is relatively constant then the 100 (1-α)% level global confidence limits may be determined from

$$\bar{e} \pm (s/\sqrt{N})t_{1-\alpha/2}(N-1)$$

where $t_{1-\alpha/2}(N-1)$ is obtained from a table of fractional points for the t distribution. If the variance definitely wanders with range then a global confidence interval is not too credible. Instead, local confidence limits for contiguous subsamples are more appropriate. For subsamples of size p the confidence limits for the k^{th} subsample are given by

$$\bar{e}_k \pm (S_k/\sqrt{p})t_{1-\alpha/2}(p-1)$$

where

$$\bar{e}_k = \sum e_n/p$$

$$s_k^2 = \sum (e_n - \bar{e}_k)^2/(p-1)$$

and the summations extend over p residuals about $\bar{e}_k$.

4.1.3 SYNTHETIC ENSEMBLES

The total collection of events (or runs) obtained during a given experiment sometimes contains events demonstrating sufficient similarity among their measurement conditions to qualify as pseudo replications. Ensembles synthesized from replications (actual or pseudo) are particularly desirable for model evaluation because they provide a statistical base that allows meaningful point-by-point comparisons.

A matrix of ensemble data is formulated as:

EVENTS (OR RUNS)

SAMPLE RANGES	n	m = 1	2	...	M
R_1	1	e_{11}	e_{12}	...	e_{1M}
R_2	2	e_{21}	e_{22}	...	e_{2M}
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
R_N	N	e_{N1}	e_{N2}	...	e_{NM}

The elements e_{nm} of the residual error matrix are formed by subtracting the prediction $\hat{L}_n$ from the measurement L_{nm} interpolated at range R_n from data recorded during event m. Mean values and mean squares are calculated in accordance with classical one-way analysis of variance (ANOVA)

procedures, where the range indices correspond to treatments or groups and the event indices correspond to replications. However, instead of testing the hypothesis of similar group means, the various mean squares are accepted as the quantities of interest, that is, local and global measures of closeness.

The range sensitive and global means are:

$$\bar{e}_{n.} = \sum_{m} e_{nm}/M \quad \text{(range sensitive)}$$

and

$$\bar{e}.. = \sum_{n} \bar{e}_{n.}/N \quad \text{(global)}.$$

The treatment mean square is given by:

$$S_T^2 = M \sum_{n} (\bar{e}_{n.} - \bar{e}..)^2/(N-1)$$

and quantifies the departure of the mean residual error obtained at each sample range from the global mean. The sequence of mean squares given by:

$$S_{n.}^2 = \sum_{m} (e_{nm} - \bar{e}_{n.})^2/(M-1) \quad (n=1,2,\ldots,N)$$

provides a set of local measures of closeness in that a distinct value is obtained for each range or range interval. These measures are not

affected by variations in distribution (or distribution parameters) with range. Rather, their credibility depends on the distributional integrity across all events at a given range. Averaging this sequence over all ranges yields the so-called residual mean square, which also serves as a global measure of closeness.

Mean square successive differences may be calculated for ensemble data in the same manner as is done for single-event data. The expressions are identical, the difference being that e_n is replaced everywhere by $\bar{e}_{n.}$, the average over events.

A global confidence interval for "ensembled" data may not be especially credible since both the mean and the variance of the residuals typically wander with range. However, if the events comprising an ensemble have been selected properly, then confidence intervals calculated for either each sample range or a select few are appropriate. The sample variance $S^2_{n.}$ is distributed as $\sigma_n{}^2\chi^2/(M-1)$ and under certain assumptions the statistic

$$\frac{\bar{e}_{n.}/(\sigma_n/\sqrt{M})}{\sqrt{S^2_{n.}/\sigma_n{}^2}}$$

has a t-distribution with M-1 degrees of freedom. As a consequence, the boundaries of a $100(1-\alpha)\%$ confidence interval are given by $e_n \pm (S_n/\sqrt{M})t_{1-\alpha/2}(M-1)$.

4.2 DISTANCE MEASURES

4.2.1 SIMPLE METRICS

An intuitive notion of a measure of closeness is some kind of distance measure. In particular the distance between measurements and predictions may be dealt with in terms of a distance function called a metric. A metric associates with each pair of points (x,y) a real number d(x,y) which satisfies certain axioms. A detailed discussion of this notion is not presented here (see for example: Royden [1968]).[8] The important point to be aware of is that a metric quantifies the degree of closeness of two points. Some common examples of metrics are:

$$d_1(L,\hat{L}) = \sum_n |L_n - \hat{L}_n| = \sum |e_n|$$

$$d_2(L,\hat{L}) = \max_n \{|L_n - \hat{L}_n|\} = \max_n \{|e_n|\}$$

$$d_3(L,\hat{L}) = \sqrt{\sum (L_n - \hat{L}_n)^2} = \sqrt{\sum e_n^2}$$

Note that once a particular metric has been chosen the closeness of predictions and measurements is expressible absolutely. Unfortunately, accuracy in the context of model evaluation must be expressible relative to something. Unless that something is universally familiar the resulting expression of accuracy is not too meaningful.

8. Royden, HL, Real Analysis, Macmillan, 1968

Consequently the productive utility of metrics as accuracy measures resides within the realm of comparative evaluation. That is, the accuracy of one model relative to that of another model can be assessed by comparing their respective metric values derived from a common set of measurements. Corresponding to a set of N sample ranges, R_1, R_2, ..., R_N, let $x = (x_1, x_2, \ldots, x_N)'$ denote a vector of measurements and let $y_1 = (y_{11}, y_{12}, \ldots, y_{1N})'$ and $y_2 = (y_{21}, y_{22}, \ldots, y_{2N})'$ denote vectors of predictions generated by model one and model two. For any metric $d(x,y)$ model one is closer to the measurements than is model two if

$$d(x,y_1) < d(x,y_2).$$

If this inequality persists regardless of which set of measurements enters into the comparison, then the results are conclusive. Otherwise the procedure is inconclusive, or, at best, is resolvable by resorting to methods of statistical inference.

4.2.2 AN ENSEMBLE MEASURE

This example is taken from the rapidly growing methodology known as pattern recognition. The particular method summarized here is described in more detail in the text by Tou and Gonzalez [1974].[9] Actually, the following measure is one of similarity in that it is sensitive to features characteristic of the data tested against.

---

9. Tou, JT and Gonzalez, RC, Pattern Recognition Principles, Addison-Wesley, 1974.

Let L_{nm} denote measured loss at range R_n for event (or run) m, and let $\bar{L}_{n.}$ denote the average over all events at range R_n, that is,

$$\bar{L}_{n.} = \frac{1}{M} \sum_{m=1}^{M} L_{nm}.$$

If $\hat{L}_n$ denotes predicted loss, then a measure of similarity between the vector of predictions $\hat{L}' = (\hat{L}_1, \hat{L}_2, \ldots, \hat{L}_N)$ and the vector of means $\bar{L}' = (\bar{L}_1, \bar{L}_2, \ldots, \bar{L}_{N.})$ is provided by

$$D^2 = (\hat{L} - \bar{L})\ C^{-1}\ (\hat{L} - \bar{L})$$

where

$$C = \begin{matrix} C_{11} & C_{12} & \cdots & C_{1N} \\ \vdots & \vdots & & \vdots \\ C_{N1} & C_{N2} & \cdots & C_{NN} \end{matrix}$$

and

$$C_{ij} = \sum_{m} (\bar{L}_{im} - \bar{L}_{i.})(L_{jm} - \bar{L}_{j.}).$$

The advantages offered by this measure over those based on sums of squares or nonparametric procedures are not readily apparent, but an obvious disadvantage is the effort required for implementation. Indeed, the problem of assembling an ensemble of events all belonging to an identifiable class is formidable in itself. Add to that the time-consuming procedure of large matrix inversion and the method loses much appeal, no matter how well it performs. On the other hand, if application of the D^2 technique is confined to range intervals of "reasonable" size, then the computational load becomes less demanding.

The D^2 technique is appropriate for determining which one of a collection of models is "closest" to an ensemble of measurements. Thus its utility in model evaluation would be limited to comparative procedures. The problem with such procedures is that the relative ranking obtained from a given application of the test is itself a random variable. Consequently, several independent applications are necessary to substantiate conclusions.

4.3 AN FOM APPROACH

The majority of methods reviewed here are based on comparing values of loss at a given range. Inverting this procedure by comparing ranges at a given value of loss is just as valid. This inverted procedure is reminiscent of figure-of-merit (FOM) analyses. For nonreverberant situations the FOM for a given set of system and target parameters is independent of range and equals the propagation loss that just satisfies the sonar equation. Corresponding to this value of loss is

the range (or set of ranges) at which target detection is just achievable. Applying this approach to propagation loss model evaluation, accuracy assessment can be based on some measure of agreement between measured and predicted ranges corresponding to prespecified values of loss.

The following techniques for implementing the "FOM approach" are suggested by concepts elementary to Lebesque integration. Let the ordinate axis of a loss-versus-range plot (transmission loss diagram) be divided into K intervals delineated by the levels, say, $y_0 < y_1 < \ldots < y_K$, where $y_0 \leq \min\{\min(L_n), \min(\hat{L}_n)\}$ and $y_K \geq \max\{\max(L_n), \max(\hat{L}_n)\}$. For the k^{th} interval let range sets E_k and $\hat{E}_k$ be defined by

$$E_k = \{R_n | y_{k-1} \leq L_n < y_k\}$$

and

$$\hat{E}_k = \{R_n | y_{k-1} \leq \hat{L}_n < y_k\} .$$

If m(E) denotes some "measure" of E, then a measure of closeness is provided by

$$\gamma = \frac{\sum_{k=1}^{K} m(E_k \cap \hat{E}_k)}{\sum_{k=1}^{K} m(E_k)} .$$

Two simple measures are suggested. Let m(e) denote the number of points in E. Then for E_k and $\hat{E}_k$ defined as above, γ is the ratio of the number of points contained in both $\cup E_k$ and $\cup \hat{E}_k$ to the number of points contained in $\cup E_k$. Since $(\cup E_k) \cap (\cup \hat{E}_k) = \cup (E_k \cap \hat{E}_k)$ and $m[\cup (E_k \cap \hat{E}_k)] \leq \Sigma m(E_k \cap \hat{E}_k)$ and $m(E_k \cap \hat{E}_k) \leq m(E_k)$ then $0 \leq \gamma \leq 1$.

As a second measure let $m^*(E)$ denote the sum of lengths of the shortest intervals containing the points in E. Then a measure of closeness, γ^*, is obtained as above but with m replaced by m^*. The measure m^* (formerly called the Lebesque outer measure, see for example Royden [1968])[8] is not quite as easily machine implementable as the counting measure, m, but it is probably more appropriate for data sets of low sample density.

5.0 STATISTICAL TEST PROCEDURES

The measures of closeness discussed in section 4.1 are quantities commonly employed in classical parameter estimation and hypothesis testing procedures. In this section those quantities and other statistics are exploited within the the framework of hypothesis testing. The discussion begins with some remarks about hypothesis tests and the assumptions necessary to implement them. Section 5.2 addresses the problem of testing one model at a time, and section 5.3 presents procedures appropriate for comparatively testing two models.

5.1 PRELIMINARY REMARKS

The statistical test procedures reviewed in sections 5.2 and 5.3 address specific hypotheses. A statistical hypothesis can be a statement about a distribution function or, more typically, one or more of its parameters. Some of the procedures reviewed here make no assumptions about distribution functions. Instead, the hypothesis addresses some characteristic of the population from which a sample is drawn.

In the framework of a null hypothesis (H_0) versus some alternative (H_i), most of the tests discussed here focus on one of the following three statements:

$H_0: \mu=0$ versus $H_1: \mu>0$,

or $H_0: \mu=0$ versus $H_2: \mu<0$,

or $H_0: \mu=0$ versus $H_3: \mu\neq 0$.

All of these statements involve a location parameter μ (eg, mean or median). In each of the first two statements H_0 is to be tested against

a one-sided location alternative (H_1 or H_2), whereas in the last one H_0 is to be tested against a two-sided location alternative (H_3). The alternatives H_1 and H_2 are preferable to H_3 as long as the location bias is suspected a priori.

To test whether or not to reject H_0 requires a test statistic, say T, which is some function of sample random variables. A decision rule is facilitated by the notion of a rejection (or critical) region, the bounds of which are called critical values. For example, to test $H_0:\mu=0$ versus $H_1:\mu>0$, a bound t_α is specified such that H_0 is rejected if $T \geq t_\alpha$. The critical value, t_α, is associated with the test's significance level, α. Specifically, t_α and α are related by a probability statement of the form

$$\alpha = Pr(t \geq t_\alpha | \mu=0) \equiv P_0(t \geq t_\alpha) ,$$

where t symbolizes all possible realizations of a random process of which T is a particular sample. Common values of α are 0.1, 0.05 and 0.01; thus the probability of incorrectly rejecting H_0 is kept small.

Most of the statistical tests discussed here are either "paired-sample" tests or "two-sample" tests. In those cases where both measurements and predictions are sampled at the same range values, a paired-sample procedure is appropriate. Paired-sample procedures are based on residual errors defined as point-wise differences between measurements and predictions. Two-sample procedures, on the other hand,

are based on the two sequences obtained after removing gross range-dependent trends from both measurements and predictions.

Let $\{L_m\}^M$, denote a sequence of M propagation loss measurements at ranges R_m, m=1,2,...,M. Similarly, let $\{\hat{L}_n\}^N$, denote a sequence of N predictions. If M = N and the sets of range values coincide, then a "paired-sample" sequence can be defined as

$$\{e_n | e_n = L_n - \hat{L}_n,\ n=1,2,\ldots,N\} .$$

For many measurement data sets, the "measured" ranges are not integer valued and uniformly spaced. Consequently, a paired-sample sequence can be formed only by means of interpolation. If for some reason interpolation is undesirable, then resort must be made to "two-sample" procedures.

For either situation, a statistical model is necessary to formulate hypotheses for testing. The simplest model assumes that the expected value of measured losses, L_m, consists of a systematic component, say f_m, and a random component, say μ_m, so that

$$L_m = f_m + \mu_m .$$

Similarly, the expected value of predicted losses, $\hat{L}_n$, consists of a deterministic component, $\hat{f}_n$, a "random" component $\hat{\mu}_n$, and a model error component, say δ, so that

$$\hat{L}_n = \hat{f}_n + \hat{\mu}_n + \delta .$$

The latter component, δ, is the parameter to be tested. The other components are assumed to be identical in the sense that $f_m = \hat{f}_m$ for coincident sample ranges and the distribution of μ is identical to the distribution of $\hat{\mu}$.

This formulation is admittedly simple inasmuch as δ must absorb any discrepancies in assumptions regarding the other two components. In fact, of course, the deterministic component is the one of primary interest, and the random component only serves to create difficulty in the matter. Neither component is likely to conform with the model precisely, and hence δ is not likely to remain constant as assumed. Nevertheless, the test objective is to determine if δ differs from zero significantly, so that modest departures from the assumed model should not seriously impair the credibility of a test for a shift in location.

There is a problem with the statistical model assumed for predictions that cannot be overlooked. For a purely deterministic model, the random component is absent for a particular event. In fact, a random component is apparent only in association with an ensemble of predictions generated in accordance with a Monte Carlo scheme (see Solomon and Merx [1974][10] and for a slightly different point of view

10. Solomon, LP and Merx, WC, Technique for Investigating the Sensitivity of Ray Theory to Small Changes in Environmental Data, J. Acoust. Soc. Am. 56:1126-1130, 1974.

see Dozier and Tappert [1978]).[11] Actually, range-dependent models can simulate, to a limited degree, some of the randomness likely to be incurred during a single measurement event. The same objection, however, does not apply to the statistical model assumed for measurements, since during a given measurement event there is randomness in both space and time. As a consequence, the significance of a test cannot be attributed entirely to model error.

Applying the assumed statistical model to paired-sample data yields a model for residual errors of the form

$$e_n = L_n - \hat{L}_n = \delta, \text{ for all } n.$$

For data that cannot be paired, two sequences are constructed: one for measurements, say

$$x_m = L_m - \tilde{L} \cong \varepsilon_m,$$

and another for predictions, say

$$y_n = \hat{L}_n - \tilde{\hat{L}} \cong \hat{\varepsilon}_n + \delta,$$

11. Dozier, LB and Tappert, FD, Statistics of Normal Mode Amplitudes in a Random Ocean. I. Theory, J. Acoust. Soc. Am. 63:353-365, 1978.

where $\tilde{L}$ and $\tilde{\tilde{L}}$ are range-trend-removal functions. Since both ε_m and $\hat{\varepsilon}_n$ are assumed to be independent and identically distributed "random" variables, any significant differences in the distributions of x and y are assumed attributable to δ. Thus, for both paired- and two-sample circumstances, the null hypothesis takes the form

$$H_o: \delta = 0,$$

and the alternative takes one of the forms

$$H_1: \delta < 0, \quad H_2: \delta > 0, \quad \text{or} \quad H_3: \delta \neq 0.$$

For conscientious objectors unwilling to accept the frailties of the assumed statistical model, the hypotheses can be stated in terms of distribution functions. For example, in the two-sample case the null hypothesis can be stated as

$$H_o: F_x(z) = F_y(z),$$

which can be tested against an alternative, say H_1, of the form

$$H_1: F_x(z) = F_y(z-\delta) \ .$$

In this way the sampled data is not explicitly broken down into systematic and random components, although the annoying issue concerning the assumed "random" character of predictions remains. The issue of "random" predictions notwithstanding, the remainder of section 5 is devoted to applying various statistical procedures to test the closeness of measurements and predictions.

5.2 ONE-MODEL TESTS

The procedures discussed in this section may be applied to statistically assess how close a particular model predicts what is observed. Again, for emphasis, the remarks of section 5.1 concerning the "random" character of predictions are reiterated. Even though random predictions can be generated using Monte Carlo methods, predictions corresponding to a particular event do not emulate the random process observed during a measurement event. Therefore, if a test infers discrepancy not all of it is necessarily attributable to model error.

The first four procedures reviewed are taken from classical statistics, wherein certain assumptions pertaining to distributions must be satisfied. The remaining procedures are nonparametric or distribution-free, and are subject to less severe restrictions.

5.2.1 REGRESSION MODEL TEST

In the following discussion a lack-of-fit test commonly applied to linear regression models (see, for example, Draper and Smith [1966])[12] is reviewed and examined for its applicability to the task of assessing the accuracy of acoustic models. For a first-order linear-regression model of the form

$$\hat{L} = \bar{L} + \beta(R-\bar{R}) \ ,$$

12. Draper, NR and Smith, H, Applied Regression analysis, Wiley, 1966.

the total sum of squares is partitioned as

$$(L_n - \bar{L})^2 = \sum(L_n - \hat{L}_n)^2 + \sum(\hat{L}_n - \bar{L})^2.$$

The sum of squares on the left-hand side accounts for variation about the mean and has N-1 degrees of freedom. The first sum of squares on the right-hand side accounts for variation "about regression," and the second accounts for variation "due to regression" and has only one degree of freedom. The sum of squares about regression when divided by the remaining degrees of freedom is referred to as the residual mean square and is denoted here by S_R^2. Thus,

$$S_R^2 = \frac{1}{N-2} \sum(L_n - \hat{L}_n)^2.$$

For an ensemble of M measurement events, each with N sample ranges ($N \geq 2M$), the mean square due to "pure" error is

$$S_e^2 = \frac{\sum_n \sum_m (L_{nm} - \bar{L}_{n\cdot})^2}{N(M-1)}.$$

The residual mean square for such an ensemble is

$$S_R^2 = \frac{\sum_m \sum_n (L_{nm} - \hat{L}_n)^2}{M(N-2)}.$$

The F statistic given by

$$F = \frac{S_R^2 - S_e^2}{S_e^2}$$

has N-2M degrees of freedom in the numerator and N(M-1) degrees of freedom in the denominator. Thus the model is rejected at the α significance level when

$$F > F_{1-\alpha}\,(N\text{-}2M,\ NM\text{-}N)\ ,$$

where $F_{1-\alpha}(N\text{-}2M,\ NM\text{-}N)$ is obtained from a table of percentage points of the F distribution.

This test can also be applied to higher-order linear regression models for which k coefficients are determined from measurement data. The only apparent change necessary to the above expressions is in the "numerator" degrees of freedom. Thus N-2M is replaced by N-M(k+1). At this stage there is strong temptation to extend the applicability of this test to nonregression models merely by adjusting the degrees of freedom. Deterministic models of propagation do not require any coefficients to be estimated from propagation measurements. Therefore, with k=0 the appropriate degrees of freedom become N-M.

Unfortunately, a change in the degrees of freedom is not the only change that occurs. The partitioning of the total sum of squares undergoes a modification as well. For linear regression models the sum of cross products, i.e., $\Sigma(L_n - \bar{L})(\hat{L}_n - \bar{L})$, vanishes as a result of

certain simplifying relationships. Similar relationships do not hold for deterministic models. The additional sum of cross products tends to compromise the integrity of the F statistic, since the residual mean square no longer accounts for all of the variation "about prediction." In fact, a deceptively small value could be generated by the F statistic, even if a large discrepancy exists between measurements and predictions. Only if the mean value of predictions equals (or is very close to) the mean value of measurements, can this test be applied with any confidence.

5.2.2 STUDENT'S T TEST

The t test is probably the most widely used test for equality of means. Let μ and σ^2 denote the population (assumed normal) mean and variance of residual errors. To test the null hypothesis $\mu=0$ against the two-sided alternative $\mu\neq 0$ at the α significance level, the t statistic, under H_0, is

$$t = \frac{\bar{e}}{S/\sqrt{N}} \quad ,$$

and the rejection region is equivalent to

$$|t| > t_{1-\alpha/2}(N-1) \quad .$$

The sample estimates $\bar{e}$ and s^2 are given by

$$\bar{e} = \frac{1}{N}\sum e_n, \quad \text{where } e_n = L_n - \hat{L}_n \quad ,$$

and

$$s^2 = \frac{1}{N-1} \sum (e_n - \bar{e})^2.$$

The appropriate value of $t_{1-\alpha/2}(N-1)$ is obtained from a table of fractional points for student's t distribution.

For many experimental situations the conditions of the central limit theorem are approximately met. As for the situation at hand, both mean and variance are likely to wander with range and the e_n are not necessarily independent for neighboring samples. However, in spite of such probable shortcomings, certain precautions can be taken to enhance the validity of the t test. Since the t distribution approaches normality (which reflects less uncertainty in s^2) as the degrees of freedom (N-1) increase, then a large sample size is desirable. On the other hand, constancy of mean and variance is more likely to sustain over contiguously grouped samples of small size. Thus the optimum situation is likely to be achieved by applying the test to large samples that preserve an acceptable degree of homogeneity.

5.2.3 CORRELATION TEST

A test commonly employed to determine the degree of relationship between two random variables is based on the sample correlation coefficient. Let x_n and y_n denote values of measured and predicted loss (with range trend removed), then the sample correlation coefficient is given by

$$r_{xy} = \frac{\sum x_n y_n}{\sqrt{\sum x_n^2 \sum y_n^2}}$$

This coefficient by itself provides a measure of closeness normalized to the range $-1<r_{xy}<1$, with $r_{xy} = \pm 1$ indicating a perfect linear relationship between x and y. Values of r_{xy} close to zero are less conclusive, in that either there could be a nonlinear relationship between x and y or the data points are simply too scattered for any relationship to be discernible. Questionable values of r_{xy} can be tested for statistical significance (see p. 413 of Brownlee [1960][13], or p. 128 of Bendat and Piersol [1971])[14] by comparing $z\sqrt{N-3}$ against $z_{\alpha/2}$, where

$$2z = \ln[(1 + r_{xn})/(1 - r_{xy})],$$

and $z_{\alpha/2}$ is obtained from tabulated values of the standardized normal distribution function. The hypothesis to be tested is that x and y are uncorrelated. Thus, if $|z\sqrt{N-3}|>z_{\alpha/2}$ then x and y are correlated (ie, not uncorrelated) at the α significance level.

13. Brownlee, KA, Statistical Theory and Methodology in Science and Engineering, Wiley, 1960.
14. Bendat, JS and Piersol, AG, Random Data Analysis and Measurement Procedures, Wiley-Interscience, 1971.

5.2.4 NONPARAMETRIC PROCEDURES

The remarks at the beginning of section 4 pertaining to "goodness of fit" testing were not intended to apply strictly to classical tests of statistical inference. They also apply to nonparametric tests as well. That is, the comparison of predictions-to-measurements or measurements-to-measurements can be formulated in terms of significance level tests based on nonparametric procedures. Details of these procedures are widely discussed in the literature (eg, Baker [1974],[15] Gibbons [1971],[16] Hajek [1969],[17] and Middleton [1969])[18] and are only briefly reviewed here.

5.2.4.1 A CHI-SQUARE TEST

The chi-square test is commonly employed in so-called "goodness-of-fit" procedures. Essentially this test compares distribution functions vis-a-vis testing for differences in certain population parameters. If a test results in rejection, the reasons for rejection are not at all specific. Thus, this test provides a measure of closeness only if the meaning of close is interpreted in a broad sense.

---

15. Baker, CR, Some Statistical Tests for the Analysis of Sonar Data, Department of Statistics, University of North Carolina, Report B-74-3, Jun 1974.
16. Gibbons, JD, Nonparametric Statistical Inference, McGraw-Hill, 1971.
17. Hajek, J, Nonparametric Statistics, Holden-Day, 1969.
18. Middleton, D, Acoustic Modeling, Simulation, and Analysis of Complex Underwater Targets II, Statistical Evaluation of Experimental Data, Applied Research Laboratories, U. of Texas at Austin, ARL-TR-69-22, 26 Jun 1969.

Let $\{x_n\}$ and $\{y_n\}$ denote sequences of size N consisting of range-detrended measurements and predictions. The sequence of predictions $\{y_n\}^N$, is divided into K categories (dB bins) and the number M_k of y_n that fall into each is determined (essentially a histogram). Similarly, let N_k denote the number of x_n that fall into the k^{th} category, then

$$x^2 = \sum_{k=1}^{K} \frac{(N_k - M_k)^2}{NM_k}$$

is approximately chi-square distributed with K-1 degrees of freedom (see for example p. 9-4 of Natrella [1966]).[19] If $x^2 > x^2_{1-\alpha}(K-1)$ the measurements are concluded to differ from the predictions at the α significance level.

5.2.4.2 KOLMOGOROV-SMIRNOV TEST

The Kolmogorov-Smirnov (K-S) test is similar to the Chi-Square test in that it tests for differences in the distributions of x and y. Again let x_n and y_n denote range-detrended measurements and predictions. In this case, however, the two sequences do not have to be of the same size.

Let $x_{(m)}$ denote the m^{th} smallest element of the sequence $\{x_m\}_1^M$.

19. Natrella, MG, Experimental Statistics, National Bureau of Standards Handbook 91, 1966.

Then

$$\{x_{(m)} | x_{(1)} < x_{(2)} < \cdots < x_{(M)}\}$$

denotes the ordered sequence of measurements. Similarly,

$$\{y_{(n)} | y_{(1)} < y_{(2)} < \cdots < y_{(N)}\}$$

denotes the ordered sequence of predictions. The corresponding sample distributions, denoted by $S_M(x)$ and $T_N(x)$, are defined by

$$S_M(x) = \begin{cases} 0 & , \quad x < x_{(1)} \\ k/M & , \quad x_{(k)} \leq x < x_{(k+1)} \\ 1 & , \quad x \geq x_{(M)} \end{cases}$$

and

$$T_N(x) = \begin{cases} 0 & , \quad x < y_{(1)} \\ k/N & , \quad y_{(k)} \leq x < y_{(k+1)} \\ 1 & , \quad x \geq y_{(N)} \end{cases}$$

From these distributions is determined the K-S statistic, D, defined by

$$D = \max_x |S_M(x) - T_N(x)| .$$

The null hypothesis of identical distributions is rejected at the significance level if

$$D \geq c ,$$

where c is obtained from tabulated values of $Pr\{D \geq c\} = \alpha/2$.

Its ease of use makes the K-S test one of the most popular nonparametric procedures available. Since sample distributions are employed, the K-S test is not sensitive to "artificially" selected categories as is the case with the chi-square test. However, like the chi-square test, it is indiscriminately sensitive to differences between the distributions being tested. That is, it is as likely to infer rejection for differences in symmetry and dispersion as it is for differences in location. What is more, the K-S test requires that the notion of randomness be attributed to predictions generated by a deterministic model - a notion that is not entirely acceptable.

5.2.4.3 PAIRED-SAMPLE TESTS

The major advantage in using nonparametric procedures based on paired-sample data is that no assumptions need be made regarding the distributions of either propagation loss measurements, L_n, or especially, of predictions, $\hat{L}_n$. Instead, the assumption is made that the pointwise differences (residual errors denoted by e_n) are independent and identically distributed over an ensemble. To clarify this assumption let $e_{nm} = L_{nm} - \hat{L}_n$, where L_{nm} denotes propagation loss measured at range R_n, $n=1,2,\ldots,N$, during event m, $m=1,2,\ldots,M$, and $\hat{L}_n$ denotes the corresponding prediction at range R_n. If $F_{nm} = Pr\{e_{nm} \leq e\}$ denotes the distribution function of residual error, then $F_{ij} = F_{kl}$ for $i=k$ but equality does not necessarily hold for $i \neq k$.

Measurements for a given event essentially constitute a time series (but with the additional complexity of spatial variation as well), and if densely sampled may have to be decimated so that the

sample spacing equals or exceeds the average decorrelation interval. Although such a procedure does not absolutely guarantee independence, it is recommended as a step preceding statistical testing.

Two nonparametric procedures for paired-sample data are discussed, both of which are easy to implement. One is called the sign test, and the other is known as the Wilcoxon signed-rank test. For each of these tests the residual errors (for a single event) are assumed to be of the form

$$e_n = L_n - \hat{L}_n = \gamma_n + \delta$$

where γ_n is a random variable with distribution unspecified and δ represents model error. The null hypothesis, $\delta=0$, is to be tested against the two-sided location alternative, $\delta\neq 0$.

The statistic for the sign test is simply the number of positive e_n. Thus

$$K = \sum_{n=1}^{N} U(e_n) \quad \text{where} \quad U(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Any zeroes ($e_n=0$) are discarded and N is reduced accordingly. Each of the 2^N possible outcomes corresponds to a Bernoulli trial with probability, say, p. Thus the distribution of K is binomial,

$$\Pr\{K \leq k\} = \sum_{n=0}^{k} \binom{N}{n} p^n (1-p)^{N-n} \quad .$$

Under H_0 $p = 1/2$ so that

$$\Pr\{K \leq k | H_0\} = \sum_{n=0}^{k} \binom{N}{n} 2^{-N} .$$

The test rejection region corresponds to K either too large or too small. Specifically, rejection occurs when $K \geq k_{\alpha/2}$ or $K \leq k'_{\alpha/2}$, where $k_{\alpha/2}$ and $k'_{\alpha/2}$ are the smallest and largest integers satisfying

$$\sum_{k=k_{\alpha/2}}^{N} \binom{N}{k} 2^{-N} \leq \frac{\alpha}{2} \quad \text{and} \quad \sum_{k=0}^{k'_{\alpha/2}} \binom{N}{k} 2^{-N} \leq \frac{\alpha}{2} .$$

Tables of the binomial distribution with p=1/2 are readily available even for N fairly large. However, the large-sample normal approximation is reportedly good for $N \geq 12$ (p. 108, Hajek [1969];[17] p. 102, Gibbons [1971]).[16] Let

$$z = \frac{K - E_0(K)}{\sqrt{\text{var}_0(K)}},$$

where $E_0(K) = N/2$ and $\text{var}_0(K) = N/4$, then the large-sample test rejection region corresponds to $|z| > z_{1-\alpha/2}$.

The Wilcoxon signed-rank statistic is given by

$$W = \sum_{n=1}^{N} r(|e_n|)U(e_n)$$

where $r(|e_n|)$ is the rank of $|e_n|$ and

$$U(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

Even though the $U(e_n)$ constitutes a Bernoulli process, the distribution of W is not quite as straightforward as that for the sign test. The α-level two-sided rejection region is equivalent to

$$W \leq \frac{N(N+1)}{2} - w_{\alpha/2} \text{ and } W \geq w_{\alpha/2},$$

where $w_{\alpha/2}$ satisfies $Pr\{W \geq w_{\alpha/2} | H_0\} = \alpha/2$. Tables are readily available (e.g., for $3 \leq N \leq 15$, see p. 269 of Hollander and Wolfe [1973]),[20] but if the residual errors are more or less symmetrically distributed about zero, then the large-sample normal approximation is reportedly good for $N \geq 15$ (p. 109, Hajek;[17] p. 113, Gibbons).[16] The large-sample test is the same as that for the sign test except the mean and the variance are given by

$$E_0(W) = \frac{N(N+1)}{4},$$

and

$$var_0(W) = \frac{N(N+1)(2N+1)}{24}.$$

20. Hollander, M, A Distribution Free Test for Parallelism, J. Amer. Stat. Assoc. 65:387-394,1970.

5.2.4.4 METHOD OF SLOPES

The method of slopes is suggested by Hollander [1970][20] as a distribution free test for the parallelism of two regression lines. The liberty is taken here of extending its applicability to the case at hand - comparing the slopes at various range points along the prediction curve with corresponding slopes calculated from measurement data. Denote propagation loss values by L_{nm}, where n corresponds to sample range, R_n, and m identifies the data set as either measured (m=1) or predicted (m=2). The sample ranges are assumed coincident for both measured and predicted data sets. For each data set, a subset of sample ranges is selected from which are formed, say, K pairs (R_i,R_j), $i \neq j$. Slope estimates are then computed from the K pairs, that is

$$b_{km} = (L_{im} - L_{jm})/(R_i - R_j),\ k=1,2,\ldots,K.$$

Care must be exercised in pairing the ranges to ensure that mutually independent slope estimates are generated for each m. From these two sets of slope estimates, slope differences are computed of the form $d_k = b_{k1} - b_{k'2}$. The subscripts k for set 1 may be selected sequentially, whereas the subscripts k' for set 2 are selected randomly. Finally, the following statistic is formed

$$W_K = \sum_{k=1}^{K} r(|d_k|)U(d_k)$$

where $r(|d_k|)$ is the rank of $|d_k|$ and $U(x) = 1$ or 0 according as $x>0$ or ≤ 0. This statistic is the Wilcoxon signed-rank test statistic discussed

in the previous section (see pp. 106-118 of Gibbons [1971]),[16] and rejects the hypothesis of equal slopes for W either too large or too small. Test implementation procedures are summarized in the previous section.

The application of this method to bottom limited situations or to many shallow water situations poses no problem, but to situations exhibiting significant structure this approach has limitations. For example, to test for equal slopes over a narrow convergence zone requires dense sampling. That is, the sample size must be large enough to support the generation of mutually independent slope estimates. To conduct a test at the 5% significance level requires at least five slope estimates. Thus, allowing contiguous but nonoverlapping sample ranges to be used in calculating the slope estimates, at least ten sample ranges are required. A minimum of ten sample points does not seem too severe, although many data sets based on air-dropped explosives would probably not qualify!

5.3 TWO-MODEL COMPARATIVE TESTS

The tests discussed in this section allow the relative performance of two models to be compared against a single measurement set. These tests offer an alternative to testing two models individually and then comparing test results. The advantages of the two-model procedures differ from one procedure to another. The first of three procedures discussed is a modified sign test which tends to mitigate effects due to large excursions in the residual errors. The second procedure contrasts selectable threshold levels, thus allowing the measure of accuracy to be

controlled. The third procedure compares two correlation coefficients. Since one coefficient can be calculated independently of the other, this procedure offers a computational advantage not available with the other two procedures.

5.3.1 MODIFIED SIGN TEST

This procedure operates on two sets of residual errors obtained from a sequence of propagation losses $\{L_n\}_{n=1}^{N}$ measured at ranges R_n, $n=1,2,\ldots,N$, and two sequences $\{\hat{L}_{mn}\}_{n=1}^{N}$, $m=1,2$, of predictions generated over the same range set by the two models being compared. The two sets of residual errors are formed by

$$e_{1n} = L_n - \hat{L}_{1n},$$

and

$$e_{2n} = L_n - \hat{L}_{2n}, \quad n=1,2,\ldots,N.$$

The usual residual computations (means and mean squares) can be executed at this stage, but the following procedure tends to mitigate effects due to large excursions in residuals. Essentially, moving averages of the absolute deviations are compared using a nonparametric approach. That is, let

$$\Phi_{mn} = \frac{1}{K-1} \sum_{k=-K/2}^{K/2} |e_{m,n+k}| , \qquad m=1,2$$

where n=K/2+1,..., N-K/2-1 (K odd). Then the statistic S defined by

$$S = \sum_{n=K/2+1}^{N-K/2-1} U(\phi_{1n} - \phi_{2n}) , \qquad n(x) = \begin{cases} 1, & x \leq 0 \\ 0, & x > 0 \end{cases}$$

can be employed to test the average relative displacements of the two sets of predictions with respect to the set of measurement data. Under the null hypothesis that the two models generate predictions that are equally close to the measurement data, S should be neither too large nor too small. The rejection region for this test is identical to the sign-test rejection region described in section 5.2.4.3.

5.3.2 MINIMUM CONTRASTS

The residual-error notation of the previous section is applicable in the following discussion. An error threshold value, say t, is specified which allows the residual errors for each model to be classified into one of two categories - pass or fail. Let p_m denote the number of e_{mn} for which $|e_{mn}| \leq t$ (pass), and let q_m denote the number that fail. A 2x2 table is formed as follows:

	Class I (pass)	Class II (fail)
model 1	p_1	q_1
model 2	p_2	q_2

From this table the smallest entry is identified as a_1, and the remaining entry within the same class is identified as a_2. The ordered pair (a_1, a_2) is known as a "contrast pair". Table A-28 of NBS Handbook 91 [Natrella, 1966][19] gives "minimum contrasts" for N=1 (1) 20 (10) 100 (50) 200 (100) 500 corresponding to α = 0.05 and 0.01 for two-sided alternatives. The ordered pairs in table A-28 are labeled (A1,A2). For the appropriate values of N and α the tabled pair (A1,A2) is found for which $A_1=a_1$. If $a_2 \geq A_2$ the two models differ with regard to the thresholded proportions considered.

5.3.3 TWO-SAMPLE CORRELATION TEST

The correlation test procedure discussed in section 5.2.3 can be extended to comparatively test the performance of two models against a given measured data set. The appropriate test statistic for this case is (p. 414 of Brownlee [1960])[13]

$$(z_1 - z_2)/[1/(N_1 - 3) + 1/(N_2 - 3)]^{1/2},$$

where allowance is made for different sample sizes. The hypothesis to be tested is that the two-sample correlation coefficients derive from the same population. Thus, the test rejection region corresponds to large values of the test statistic.

Since the z_i are calculated independently without regard for sample size, this procedure can be employed to "remotely" compare models.

For example, if the correlation coefficient r_{xy_1} (calculated at facility A) between model y_1 and data set x is known along with the sample size N_1, and if the same data set x (or some subset) is accessible at facility B, then r_{xy_2} between model y_2 and data set x can be calculated and the two-coefficient test statistic can be formed as indicated above. Such a procedure might be worthwhile adopting for the initial check-out stages of a new model. That is, the performance of a new model against a given measured data set could be "statistically" compared with the performance of an established model against the same data set, and without the need to execute both models on the same computer. However, at the initial check-out stage, benchmark tests against closed-form solutions are preferable.

6.0 CONCLUDING REMARKS

The preceding sections briefly summarize the current status of acoustic model evaluation, with special attention given to accuracy assessment methods as applied to propagation models. An attempt is made in section 3 to capture and relate to the reader the essential attributes of the methodologies developed under POSSM and MEP. The discussions presented in sections 4 and 5 are intended to demonstrate how standard quantities and procedures from both classical and nonparametric statistics can be applied to "time"-series data exhibiting a trend. The various moments, metrics and test procedures discussed are readily available in "statistical" software packages at most computer centers.

Model evaluation procedures as reviewed here are primarily intended for "automated" implementation. That is, some sacrifices in "analytical" or "interpretive" considerations are made to allow "wholesale" comparative evaluation of many candidate models against a variety of measured data sets. As an example of an interpretive model evaluation process, the reader is referred to the report by Hanna [1975].[21]

For the most part the collective procedures of POSSM and MEP appear to provide an adequate repertoire of techniques for evaluating propagation models and for statistically analyzing measured data sets.

---

21. Hanna, JS, An Example of Acoustic Model Evaluation and Data Interpretation, Acoustic Environmental Support Detachment TN-75-08, Dec 1975.

These same techniques, with little or no modification, should be applicable to reverberation prediction-measurement assessments as well. Their applicability to the problem of evaluating noise models, however, may be another matter.

The status of MEP software as reflected in the memorandum by Stieglitz [1974][7] indicates a means is available to perform error analyses in a routine manner. Even though wholesale application of error analysis schemes may not be practical, the introduction of such schemes to the evaluation of models against complete measured data sets is desirable. The error analysis methodology initiated by Cavanaugh [1974][6] needs to be complemented with replicated data sets. Without the support of a statistically adequate data base, error analysis procedures cannot be implemented.

Replicated data sets are also valuable for less elaborate techniques. For example, an estimate of pure error can be subtracted from a given measurement event to yield "nonrandom" components. Differences between model predictions and "derandomized" measurements can then be quantified using the methods of section 4.

As a final note the measures and procedures discussed in sections 4 and 5 are summarized below.

SINGLE-EVENT MEASURES

$\bar{e} = \frac{1}{N} \sum e_n = \frac{1}{N} \sum (L_n - \hat{L}_n)$ — Mean deviation of predictions ($\hat{L}_n$) from measurements (L_n)

$s^2 = \frac{1}{N-1} \sum (e_n - \bar{e})^2$ — Total mean square deviation

$d^2 = \frac{1}{N-1} \sum (e_{n+1} - e_n)^2$ — Mean square successive differences

$s = \sqrt{\frac{1}{N} \sum e_n^2}$ — RMS error

METRICS

$d_1 = \sum |e_n|$ $\qquad$ $d_2 = \max\{|e_n|\}$ $\qquad$ $d_3 = \sqrt{\sum e_n^2}$

ENSEMBLE MEASURES

$\bar{e}_{n.} = \frac{1}{M} \sum_m e_{nm}$ — Ensemble mean - the mean error over M events at range R_n

$\bar{e}_{n..} = \frac{1}{N} \sum_n \bar{e}_{n.}$ — Grand mean

$S_T^2 = \frac{M}{N-1} \sum_n (\bar{e}_{n.} - \bar{e}_{..})^2$ — Mean square deviation of ensemble mean from grand mean

$S_{n.}^2 = \frac{M}{M-1} \sum_m (\bar{e}_{nm} - \bar{e}_{n.})^2$ — Mean square deviation of errors from ensemble mean

$S_R^2 = \frac{1}{N-1} \sum_n S_{n.}^2$ — Residual mean square

$$d_T^2 = \frac{1}{N-1} \sum_n (\bar{e}_{n+1,.} - e_{n.})^2$$

Mean square successive difference of ensemble means

$$D^2 = (\hat{L} - \bar{L})' C^{-1} (\hat{L} - \bar{L})$$

Distance measure between vector of predictions ($\hat{L}$) and vector of means ($\bar{L}$)

STATISTICAL TESTS

CLASSICAL

- Regression Model
- Student's t
- Correlation

NONPARAMETRIC

- Chi-square
- Kolmogorov-Smirnov
- Sign
- Wilcoxon

TWO-MODEL TESTS

- Modified Sign
- Minimum Contrasts
- Two-Sample Correlation

As a final reminder, all two-sample tests require that a "random" component be assumed for predictions as well as for measurements. Since such an assumption is not especially credible, only those procedures based on residual errors (paired samples) are recommended for model evaluation. If the residuals are approximately normal, then the classical student's t test is appropriate; otherwise a nonparametric procedure (eg, Sign or Wilcoxon) is preferred.

REFERENCES

1. Forman, L, Comparative Evaluation Methods for Propagation Loss Model, Computer Sciences Corporation Unpublished Report, Jul 1975.

2. Lauer, RB and Skory, J, The Quantitative Comparison of Model Outputs with Experimental Data - A STAMP Program Application, Naval Underwater Systems Center TM TA11-46-75, 15 Jul 1975.

3. DiNapoli, FR, Computer Models for Underwater Sound Propagation, Naval Underwater Systems Center TD 5276, 31 Oct 1975.

4. Lauer, RB and Sussman, B, A Methodology for the Comparison of Models for Sonar Systems Applications, Volume I, Naval Sea Systems Command Report SEA 06H1/036-EVA/MOST-10, 9 Dec 1976.

5. Lauer, RB and Sussman, B, A Methodology for the Comparison of Models for Sonar System Applications, Volume II, Naval Sea Systems Command Report SEA06H1/036-EVA/MOST-11 (to be released).

6. Cavanaugh, RC, Transmission Loss Model Evaluation Package, Part I: The Approach, Acoustic Environmental Support Detachment (unpublished report), 1 May 1974.

7. Stieglitz, R, Informal Documentation-Model Evaluation Package, Acoustic Environmental Support Detachment memo AESD:RS:dl of 28 Jun 1974.

8. Royden, HL, Real Analysis, Macmillan, 1968.

9. Tou, JT and Gonzalez, RC, Pattern Recognition Principles, Addison-Wesley, 1974.

10. Solomon, LP and Merx, WC, Technique for Investigating the Sensitivity of Ray Theory to Small Changes in Environmental Data, J. Acoust. Soc. Am. 56:1126-1130, 1974.

11. Dozier, LB and Tappert, FD, Statistics of Normal Mode Amplitudes in a Random Ocean. I. Theory, J. Acoust. Soc. Am. 63:353-365, 1978.

12. Draper, NR and Smith, H., Applied Regression Analysis, Wiley, 1966.

13. Brownlee, KA, Statistical Theory and Methodology in Science and Engineering, Wiley, 1960.

14. Bendat, JS and Piersol, AG, Random Data Analysis and Measurement Procedures, Wiley-Interscience, 1971.

15. Baker, CR, Some Statistical Tests for the Analysis of Sonar Data, Department of Statistics, University of North Carolina, Report B-74-3, Jun 1974.

16. Gibbons, JD, Nonparametric Statistical Inference, McGraw-Hill, 1971.

17. Hajek, J, Nonparametric Statistics, Holden-Day, 1969.

18. Middleton, D, Acoustic Modeling, Simulation, and Analysis of Complex Underwater Targets II, Statistical Evaluation of Experimental Data, Applied Research Laboratories, U. of Texas at Austin, ARL-TR-69-22, 26 Jun 1969.

19. Natrella, MG, Experimental Statistics, National Bureau of Standards Handbook 91, 1966.

20. Hollander, M, A Distribution Free Test for Parallelism, J. Amer. Stat. Assoc. 65:387-394, 1970.

21. Hanna, JS, An Example of Acoustic Model Evaluation and Data Interpretation, Acoustic Environmental Support Detachment TN-75-08, Dec 1975.